The Relation between Teachers' Judgment Accuracy of Students' Performances in Reading Comprehension and Adaptive Scaffolding

Sylvia Kasperink

Utrecht University

Abstract

Educational practices are adapted to teachers' judgments of their students' performances, so it is important that teachers make accurate judgments. The author examined if teachers' judgment accuracy could predict adaptive scaffolding, an educational practice. Measuring teachers' judgment accuracy differed between teachers' judgments based on three conditions: student names (1), student names in combination with by students generated cause-and-effect relations in diagrams (2), and cause-and-effect relations (3). Students from the ninth grade of secondary school and their teachers were included. Results show that teachers' judgment accuracy does not differ between the three conditions and that teachers' judgment accuracy does not predict adaptive scaffolding.

Keywords: teachers, judgments, accuracy, adaptive scaffolding, students

The Relation between Teachers' Judgment Accuracy of Students' Performances in Reading Comprehension and Adaptive Scaffolding

Teachers' judgments on knowledge, attitudes, and behavior of their students have implications for educational practices (Gaines & Davis, 1990; Shavelson & Stern, 1981). With teachers' judgments is meant that teachers make estimations of students' performances on a test in advance. Educational practices, as instructional techniques, pace, and support are adjusted on teachers' judgments (Feinberg & Shapiro, 2003; Gaines & Davis, 1990; Hoge & Coladarci, 1989). Teachers judge the degree of support—scaffolding—students need, within the Zone of Proximal Development (ZPD) (Wood, Bruner, & Ross, 1976). The ZPD is the difference between the actual developmental studentlevel and the maximal level a student could reach with support (Vygotsky, 1978).

Limited research has focused on the relation between teachers' judgments and educational practices. Research suggest that teachers must make accurate judgments on students' learning to effectively adjust instruction to facilitate learning (Helmke & Schrader, 1987; Thiede et al., 2015). This study extends literature about teachers' judgments and educational practices by examining if teachers' judgment accuracy (TJA) on students' performances could predict adaptive scaffolding.

Teachers' support can be mentioned as scaffolding, if teachers adapt their control to the understanding of the students (see Contingency of instruction, a condition for scaffolding) (Wood, Wood, and Middleton, 1978). For this, teachers should know what their students understand at school. Because of this, the current study focuses on students' comprehensions, specifically reading comprehension.

Because Thiede et al. (2015) found that when TJA is low, teachers are not able to identify students' instructional needs, this study attaches importance to measuring TJA. TJA is the degree in which teachers' judgments correspond with students' test performances, without influencing the performances (Hoge & Coladarci, 1989). TJA is dependent of two relationships: cue utilization and cue diagnosticity (see fig. 1) (Brunswik, 1956; Pyc, Rawson, & Aschenbrenner, 2014). Teachers use cues to make judgments—called cue utilization. Cue diagnosticity is the degree to which a cue is predictive for students' test performances (Pyc et al., 2014). The more diagnostic a cue, the higher the TJA (Koriat, 1997).

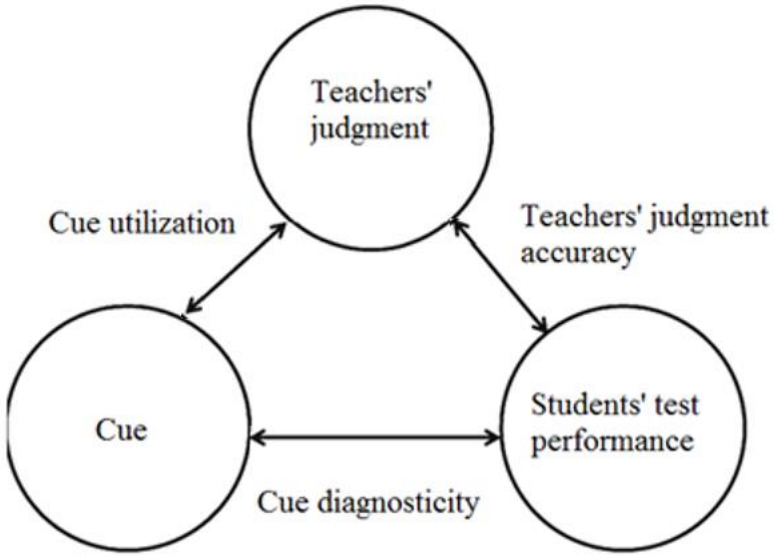

Figure 1. Relations between cues, judgments, and test performance (Dunlosky & Thiede, 2013).

This study extends literature on TJA—especially in the field of reading—by investigating TJA based on the student-related and the content-related judgment perspective from Hoth, Döhrmann, Kaiser, Busse, König, and Blömeke (2016) (see Teachers' judgments: Cue utilization, cue diagnosticity, and implications for educational practices).

Judgments based on cause-and-effect relations—the content-related perspective—are included in the study, because previous research reveals that generating cause-and-effect relations cause better text comprehension (Butcher, 2006; Gobert & Clement, 1999; McCrudden, Schraw, Lehman, & Poliquin, 2007), and better text comprehension is associated with monitoring accuracy (Anderson & Thiede, 2007; Van Loon, De bruin, Van Gog, Van Merriënboer, & Dunlosky, 2014). It is possible that better text comprehension by cause-and-effect relations is also associated with TJA.

Respectively, the theoretical background on the research questions is described. Firstly, the focus will be on the ZPD and the within supplied adaptive scaffolding. Subsequently, a condition for scaffolding, contingency of instruction, is discussed. Then, previous research on TJA in the field of reading is shown. Finally, research on teachers' judgments is discussed; teachers' use of cues, cue diagnosticity, and its implications for educational practices. This subject is divided in a student-related and a content-related perspective on judgments, with the latter focusing on cause-and-effect relations in diagrams.

Zone of Proximal Development and Adaptive Scaffolding

The study´s focus is on the relation between TJA and the within the ZDP supplied scaffolding (Fernández, Wegerif, Mercer, & Rojas-Drummond, 2001; Van de Pol & Elbers, 2013), specifically adaptive scaffolding. Adaptive scaffolding means that teachers judge individual students' need for support, provide this support, and fade the support when students' self-confidence and abilities increase. Adaptive scaffolding differs from fixed scaffolding, in which support is determined in advance and is valid for all students (Molenaar & Roda, 2008).

Contingency of Instruction, a Condition for Scaffolding. For adaptive scaffolding, contingency of instruction must be present. This means that teachers' control is adapted to the level and understanding of the individual students. Control should increase when the students make mistakes and decrease when the students' abilities increase (Van de Pol & Elbers, 2013). If the students master the task, control is fully transferred to the students (Molenaar & Roda, 2008).

Four subtypes of support show the relation between contingency, challenge, and student learning. Two of these subtypes contain non-contingent support (NA). Here, the degree of control is not adjusted to the students' initial understanding. The first subtype non-contingent support is increasing control, when students' initial understanding is good, teachers underestimate students' understanding, and the degree of challenge is little, so there is no learning (NA+). The second subtype non-contingent support is decreasing control, when students' initial understanding is poor, teachers overestimate students' understanding, and the challenge is too much, so there is no learning (NA-). The remaining two subtypes include contingent support (A). Here, the degree of control is adjusted to the students' initial understanding. The first subtype contingent support is increasing control, when students' initial understanding is poor, teachers correctly estimate students' understanding, and the challenge is appropriate, so learning takes place (A+). The second subtype contingent support is decreasing control, when students' initial understanding is good, teachers correctly estimate students' understanding, and the challenge is appropriate, so learning takes place (A-) (Van de Pol & Elbers, 2013).

Teachers' Judgment Accuracy in the Field of Reading

To effectively give instructions to students—an educational practice—teachers must make accurate judgments (Thiede et al., 2015). To examine if this association is valid in the relation between TJA and scaffolding—another educational practice—is it prerequisite that TJA is examined. In the field of reading, research indicated a strong association between teachers' judgments and students' performances. Demaray and Elliott (1998) found a significant correlation of .82 between teachers' judgments and students' reading scores on the Kaufman Test of Educational Achievement. In addition, teachers accurately predicted students' performances on reading comprehension tests (Feinberg & Shapiro, 2009).

Teachers' Judgments: Cue Utilization, Cue Diagnosticity, and Implications for Educational Practices

Teachers could use a student-related perspective, a content-related perspective, or both perspectives to judge students' performances (Hoth et al., 2016). These perspectives are associated with different use of cues, which differ in diagnosticity and anticipations on teacher strategies.

Student-related perspective. When teachers use a student-related perspective, they base their judgments on students' knowledge, attitudes, behavior, and understanding on a large amount of information—cues—about their students, which are derived from multiple resources (Alvidrez & Weinstein, 1999; Hoth et al., 2016; Kaiser, Möller, Helm, & Kunter, 2015; Shavelson & Borko, 1979). Examples of these cues are students' intelligence, motivation, topic interest, academic self-concept, and socioeconomic status (SES). These cues are among others based on teachers' own informal observations, other teachers' anecdotes, students' standardized test scores, and students' school performances (Shavelson & Borko, 1979). When teachers know their students and could only base their judgments on

student names—use the student-related perspective—they could use the above mentioned cues. Cues differ in diagnosticity. High diagnostic cues are students' motivation, interest, and academic self-concept. Motivation consists of implicit beliefs of intelligence, self-efficacy, and approach achievement goals. If students belief intelligence could grow (incremental belief), they could successfully perform a given task (high-self-efficacy), and belief task mastery is more important than showing performance to others (mastery approach goal), their academic motivation is high. Academic motivation predicts academic achievement, $r = .32$ (Mega, Ronconi, & De Beni, 2014). Schiefele (1992) found that interest in a topic causes meaningful learning, long term knowledge storage, and good text comprehension ($p < .001$), and the higher the interest, the higher the performances on tests about texts. Academic self-concept is positively associated with students' academic achievement (Guay, Marsh, & Boivin, 2003; Seaton, Parker, Warsh, Craven, & Yeung, 2013). However, Kaiser et al. (2015) found that self-concept is not a diagnostic cue for achievement. A possible explanation for this conflicting finding is that Kaiser et al. (2015) focused on mathematical achievement and the other researchers on general academic achievement. SES is a low diagnostic cue. SES include parents' income and educational level and has a positive, but weak correlation with students' academic achievement for ninth grade students ($r = .176$) (White, 1982).

With regard to educational practices, Hoth et al. (2016) suggest that when teachers use a student-related perspective, they effectively anticipate their teaching methods and instructional organization on these students.

Content-related perspective. A content-related perspective means that teachers make judgments on cues drawn from the content of a subject (Hoth et al., 2016). Hoth et al. (2016) examined a content-related mathematical perspective for mathematical content. The current

study examines a content-related reading comprehension perspective, in which teachers could base their judgments on by students' generated cause-and-effect relations from texts that they have read.

Cause-and-effect relations in diagrams. In the current study, TJA is among others based on cause-and-effect relations. Cause-and-effect relations can be effectively represented in diagrams (see fig. 2 for an example). These causal diagrams facilitate three types of effects: direct effects, indirect effect (serial effects)—by a mediating variable—and multiple effects that occur at the same time (McCrudden et al., 2007). Providing causal diagrams—besides a presented text—facilitates learning when the text's structure is simplified by the diagram and cause increased mental models and understanding of a text (Butcher, 2006; Gobert & Clement, 1999; McCrudden et al., 2007).

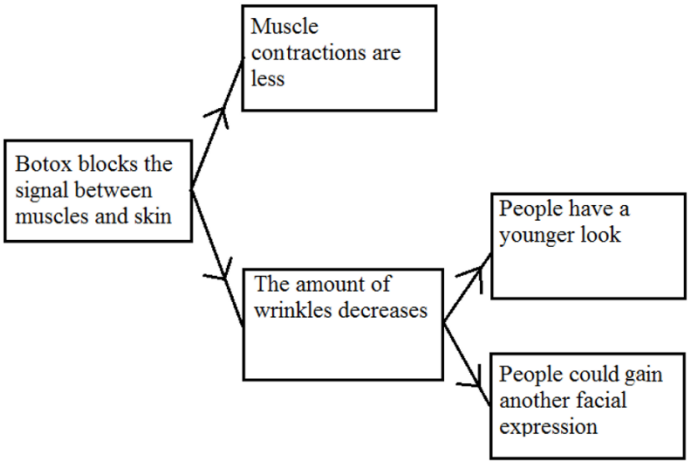

Figure 2. A completed diagram for the text 'Using Botox' that students in the current study are asked to read. The arrows represent 'therefore' (Van Loon et al., 2014).

Cue utilization, cue diagnosticity, and educational practices with regard to cause-and-effect relations in diagrams. When teachers judge students' performances based on cause-and-effect relations, they could use cues from the content-related reading perspective.

No research has focused on cues teachers draw from cause-and-effect relations in diagrams for judging. Most research on cue utilization and diagnosticity has focused on monitoring accuracy and cues drawn by students. For example, Van Loon et al. (2014) found in a study about monitoring accuracy that when students generated cause-and-effect relations in diagrams, they used cues for predicting their test performances. These cues—whether or not filled in diagram boxes—were diagnostic for their test performance. It is possible that when teachers base their judgments on these cues, they are also diagnostic for test performances.

The Current Study

In the current study it was investigated if TJA predicted adaptive scaffolding. By doing this, I examined the predictive value of TJA based on two conditions on adaptive scaffolding. These conditions differed in TJA based on student names (condition 1) and on student names in combination with cause-and-effect relations (condition 2).

For examining this relation, TJA was examined based on one more condition; by students generated cause-and-effect relations (condition 3). The underlying reason for not including this condition in research question two is explained in the method section. These three conditions align the two judgment perspectives, and the combination of them from Hoth et al. (2016). The main research questions are:

1. To what degree is there a difference between teachers' judgment accuracy on students' performance on reading comprehension, based on student names (condition 1), on student names in combination with cause-and-effect relations (condition 2), and cause-and-effect relations only (condition 3)?
2. Can adaptive scaffolding be predicted by TJA?

a. Can adaptive scaffolding be predicted by TJA based on student names (condition 1)?
b. Can adaptive scaffolding be predicted by TJA based on student names in combination with cause-and-effect relations (condition 2)?

With regard to research question one, teachers made judgments on the degree of how students would perform on reading comprehension tests. Teachers based one third of their judgments on student names, one third on student names in combination with by students generated cause-and-effect relations, and one third only on cause-and-effect relations. Teachers make use of a large amount of cues in their judgments (Borko & Niles, 1982; Brunswik, 1956; Shavelson & Borko, 1979). Because of the limited research on cue diagnosticity of cues used in judging cause-and-effect relations, it is hard to compare cue diagnosticity of this condition with cue diagnosticity of cues used for judgments based on student names. For this reason, the hypothesis regarding which judgment condition (based on student names or cause-and-effect relations) is associated with the highest TJA, was not based on cue diagnosticity. This hypothesis was based on previous research on associations. Thiede et al. (2015) and Helmke and Schrader (1987) suggest that the higher the TJA, the more effective teachers adjust instruction to facilitate learning. Hoth et al. (2016) suggest that when teachers use a student-related perspective—base their judgments on student names, they adjust their instruction to these students. Because of this, I expect that when teachers' judgments are based on student names—the student-related perspective—TJA will be higher than when teachers' judgments are based on cause-and-effect relations (hypothesis one).

However, for judgments based on student names in combination with cause-and-effect relations, the hypothesis can be based on previous literature on cue diagnosticity; the

higher cue diagnosticity, the higher the TJA (Koriat, 1997). When teachers make judgments based on the combination, they could use the cues from student names and cause-and-effect-relations, which causes the highest cue diagnosticity of the three judgment conditions. Because of this, I expect that when teachers make judgements based on student names in combination with cause-and-effect relations, this is associated with the highest TJA among the three conditions (hypothesis two).

With regard to research question two, students completed questionnaires which were related to (perceived) adaptive scaffolding. Because the higher the TJA, the more effective teachers adjust their instruction to facilitate learning (Thiede et al., 2015; Helmke and Schrader, 1987), I expect that both TJA based on student names predict adaptive scaffolding (hypothesis three) and TJA based on student names in combination with cause-and-effect relations predict adaptive scaffolding (hypothesis four). Note that research question two is based on teachers' judgments on students' performance on reading comprehension.

The research has practical interests. If the results show that TJA based on student names in combination with cause-and-effect relations has a significant added value on TJA in comparison to TJA based on student names, and that the higher the TJA, the higher the degree of scaffolding, the educational practice should focus more on teaching students how to interpret and generate cause-and-effect relations in diagrams (Demaray & Elliot, 1998) and explaining teachers to use these cause-and-effect relations to make judgments. With adaptive scaffolding, students do not passively listen to information, but are asked to build on their already stored information to store new information. Also, the students will receive positive feedback. Especially for students with low self-concept and learning problems, positive feedback enhances motivation (Van der Stuyf, 2002).

Method

Participants

Teachers (10 women, 5 men, M_{age} = 40.40, age range: 24 - 58 years) from five secondary schools in the Netherlands were recruited by team leaders with recruitment letters. Teachers were included if they taught ninth grade students. Their students had to follow the third year of higher general secondary education (HAVO; five-year duration) or the third year of pre-university secondary education (VWO; six-year duration). They taught the following subjects: geography (N = 1), biology (N = 1), English (N = 4), history (N = 3), Dutch (N = 4), and/or chemistry (N = 2). Nine teachers taught HAVO students and six teachers taught VWO students. The teachers were tested at school in absence of their students.

After recruiting teachers, their HAVO or VWO students (N = 402) from the ninth grade of five secondary schools were recruited with recruitment letters. Three students did not give permission and were excluded from the study. Also, each student received a letter with information about the purpose of the research and the researchers' contact information for their parents/caregivers, whom they could contact if they did not want to give permission for using their child's test data. Three parents did not give permission and 20 students were absent. In total N = 376 students participated in the study. Students were randomly assigned to groups who made the diagram or 'filler' task (two thirds diagram task, one third 'filler task'). In advance of the experiment, six students of the group who filled in diagrams and six students of the group who made the 'filler task' were randomly chosen as 'focus-students' for the teachers' part of the experiment, without knowing so. In total there were N = 266 focus-students, whose age is known from N = 235 students (M_{age} = 15.5, age range: 12.00 - 16.83 years) and gender is known from N = 237 (142 girls, 95 boys).

Materials and Procedure

The materials used in the study were divided in materials for students and teachers. The quality of the materials was enhanced by practicing both the experimenter and the participant of the materials together. The experiment for the students existed of five stages: pre-experiment instructions and practicing, text reading, diagram or 'filler tasks', test, and questionnaires. The first four stages of the experiment are based on the method used in Van Loon et al. (2014). The teachers also followed the stage pre-experiment instructions and practicing for students. After this stage, they followed the following three stages: filling in general information and reading texts, pre-experiment instruction and practicing, and making judgments.

Pre-experiment instructions and practicing (students). In advance of the experiment, students were told by the experimenter that they would read six texts, and then some of them would make diagrams; the others would make 'filler tasks', and that finally all the students would answer questions about the texts. The students were asked to read example text one— 'Sporting is healthy'—and were given examples of questions about cause-and-effect relations. They learned that cause-and-effect relations could be represented in diagrams, so they got used to cause-and-effect relations in diagrams. Following, the students were asked to read example text two— 'Suburbs'—and to draw a cause-and-effect diagram on their own, which suited the text. Finally, the experimenter explained them distinct cause-and-effect relations, namely cause-and-effect relations in which events could occur simultaneously or serially.

Text reading. Students were asked to read the texts: 'Metro wagons sunk', 'Using Botox', 'The Suez Canal', 'Music makes smarter', 'Money does not make happy', and

'Renovation of concrete buildings'. These texts were chosen based on a pilot study conducted by Van Loon et al. (2014), which revealed that these texts had an adequate degree of difficulty; the average achievement on tests about cause-and-effect relations was 40.83% (SD = 24.34) and 82.5% of the students could represent cause-and-effect relations of the text in diagrams (SD = 24.05). The text had an average number of 171.27 words, ranging from 162 – 189 (Van Loon et al., 2014). There were six versions handed out, in which the six texts were offered in different sequences by Latin Square Design. For examples of two texts, see Appendix A.

Diagram or 'filler task'. Because teachers have to make judgments based on three different conditions, two thirds of the students filled in diagrams, and one third of the students made a 'filler task'. The students who were asked to fill in diagrams, received paper books with blank diagrams, with only one part of each diagram for each texts filled in (see fig. 3 for an example). The students who were asked to make a 'filler task', compared two figures and encircled the differences (see fig. 4 for an example).

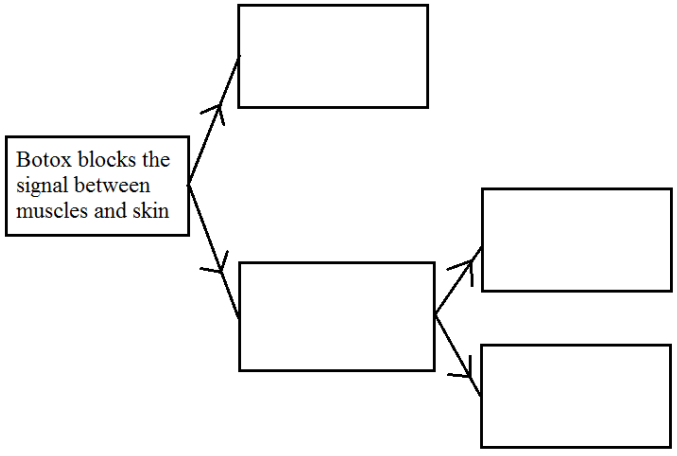

Figure 3. An empty diagram for the text 'Using Botox' that students were asked to fill in. The arrows represent 'therefore'.

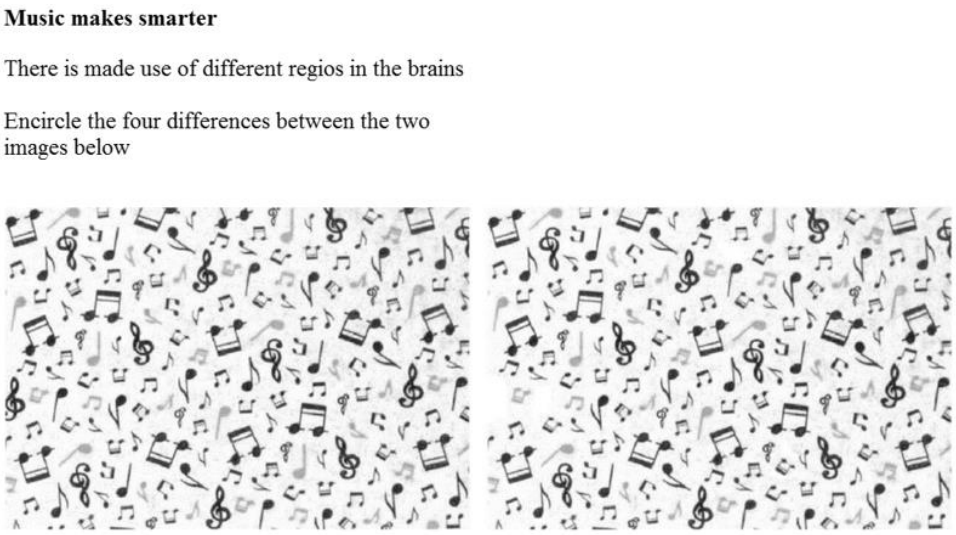

Figure 4. An example of a 'filler task' for the text 'Music makes smarter', which the students were asked to read.

Test. All students were asked to take tests about the six texts. For each text the students were asked to answer a question in which they had to fill in four cause-and-effect relations. The questions were represented in the same order as the texts and the students were not allowed to read the texts again or make use of the diagrams.

Questionnaires adaptive scaffolding. Adaptive scaffolding is in the current study defined as the by students' perceived teachers' ability to judge the students' need for support, provided support, and fading of support when students' abilities and self-confidence increase. The students were asked to fill in two questionnaires concerning adaptive scaffolding, which explicitly focused on by individual students' perceived adaptive scaffolding. Before filling in the questionnaires, students were told which teacher the questions concerned, and that the questions concerned received support when they worked independently. The experimenter and the students made one example question together. Conducting two questionnaires that measure the same construct causes methodological triangulation (Baarda, 2009).

***Perceived adaptivity questionnaire*.** The Perceived Adaptivity questionnaire (PAQ) for students (Van de Pol, Kroonenberg, Poorthuis, & Mainhardt, in progress) is used. This questionnaire consisted 27 items, divided on the four subscales/types that show the relation between contingency, challenge, and student learning (Van de Pol & Elbers, 2013) (see Contingency of instruction, a condition for scaffolding). Subscale NA+ consists of six items; $\alpha = .74$; subscale NA- consists of seven items; $\alpha = .84$; subscale A+ consists of six items; $\alpha = .88$; and subscale A- consists of eight items; $\alpha = .80$. A Cronbach's alpha score of $.90 > \alpha > .80$ indicates a good internal consistency; $.80 > \alpha > .70$ indicates acceptable internal consistency (George & Mallery, 2003). Students were asked to respond to the items on a 5-point Likert-type scale (ranging 1 = *is not correct* and 5 = *is correct*). An example of an item (NA+): 'This teacher helps me with things I already understand'.

***Adaptive intervention*.** The Adaptive Intervention questionnaire (AIQ) (Bürgermeister et al., 2011) is used. This questionnaire has no subscales and consists of six items; $\alpha = .86$. An example of an item: 'I had the feeling this teacher understood the difficulties I had when working on assignments'. Students were asked to respond to the items on a 4-point Likert-type scale ranging from 1 to 4. The first three questions ranged from 1 (*almost never*) to 4 (*almost every lesson*) and the latter six questions ranged from 1 (*totally not right*) to 4 (*totally right*).

Filling in general information and reading texts. Teachers were asked to fill in general information about themselves, for example name, teaching experience, and subject(s) taught. Besides this, they were asked to read the six texts that students also had to read.

Pre-experiment instruction and practicing (teachers). Teachers were instructed that they had to predict the performance of a sample of their own students on the test. They

were told that they had to make predictions of performances based on student names, on student names in combination with filled in diagrams, and on filled in diagrams. Teachers were asked to encircle the amount of rightly answered questions about cause-and-effect relations for each student on a range from zero (*zero rightly answered questions*) to four (*four rightly answered questions*). Finally, they practiced together with the experimenter how to make predictions (judgments).

Making judgments. Teachers were asked to make judgments on each of their 12 focus students' performances on each of the six texts. It is important to note that the diagrams for the judgment condition cause-and-effect relations were obtained from six students from previous research (Van Loon et al., 2014). This is because otherwise teachers would make judgments based on cause-and-effect relations from their own students and might relate these cause-and-effect relations to particular students by handwriting recognition. For each student, six figures were presented in which they could encircle the amount of right answers. Teachers made judgments based on student names, student names in combination with cause-and-effect relations, and cause-and-effect relations. The focus-students who were judged based on names only, were obtained from both students who made 'filler tasks' and diagrams.

Scoring

Test scoring. The amount of correct cause-and-effect relations was scored on a range from 0 to 4 as per McCrudden et al. (2007) and Van Loon et al. (2014). Students' answers were also correct when they were not literally the same as the response model—not verbatim, but a paraphrase—so they showed they understood the meaning of the texts, because text comprehension is measured (in line with Diakidoy, Mouskounti, & Ioannides, 2011; McCrudden et al, 2007; Leopold & Leutner, 2012). Also it was noted which relations

students have made and haven't made. Independently, different raters scored 10% of the diagrams, whose inter-rater reliability was 61.7% (Kappa = .62), indicating a good inter-rater reliability (Cohen, 1988).

Design

The current study has a quantitative research design. Efforts were made to examine a large amount of students and teachers from the ninth grades of HAVO and VWO, so the conclusions drawn are as representative as possible for these students. Moreover, a quantitative design caused more objectivity, because no subjective interpretations of the data were made. The study has a within-subject design, because teachers made judgments based on all three different conditions (student names, cause-and-effect relations and the combination of them), but themselves were not divided in conditions.

Analysis

For analysis, SPSS version 23 is used. For testing the hypotheses, relative teachers' judgment accuracy (TJA) had to be measured. In the current study, TJA is defined as the association between teachers' judgments and students' performances in reading comprehension. I used Goodman-Kruskal gamma correlations—a rank correlation—to measure relative TJA. This is a non-parametric test for ordinal variables and has been used in other studies for measuring accuracy (Van Loon et al., 2014; Maki & Serra, 1992; Thiede, Anderson, & Therriault, 2003). The value of gamma indicates how strong the association is between teachers' judgments and students' performances, with values ranging from -1 (*negative association*) to +1 (*positive association*). A gamma of 0 reflects no association between teachers' judgments and students' performances. For each focus-student, intra-individual calculations were made for measuring TJA, by measuring the gamma correlations

between teachers' judgments and the focus-students' performances. In this way, students were ranked on the strength of the association. The non-focused-students' performances were not included in the analysis.

Hypotheses one and two were tested by a one-way between groups ANOVA. This analysis tests for significant differences between dependent means of the judgment conditions (on student names only, on student names in combination with by students generated cause-and-effect relations, or only on cause-and-effect relations).

Hypotheses three and four were tested by linear regression analyses. Before conducting this analysis, the scores on the questions 3, 7, 8, 10, 12, 14, 15, 16, 19, 21, 23, 24, and 26 from the PAQ and the scores on question 7 from the AIQ had to be reversed. By doing this, the higher the score on both questionnaires, the higher the score on adaptive scaffolding. Because the maximum achievable score on both questionnaires differed from each other, both questionnaire scores were transformed to the same scale, so both questionnaires received the same scale. The scores on the PAQ were multiplied by .74 and the scores on the AIQ were multiplied by 2.78, so both scores were indicated on a scale of 100. Mean scores were compared between the PAQ (M = 71.39, SD = 8.65) and the AIQ (M = 67.53, SD =15.63), and did significantly differ from each other, $p < 001$. For this reason, several linear regression analyses were executed with different operationalization of the dependent variable adaptive scaffolding. The first dependent variable adaptive scaffolding (PAQ) was only measured by the PAQ; the second dependent variable adaptive scaffolding (AIQ) was only measured by the AIQ; the third variable adaptive scaffolding (mean) consisted of the mean of the scores on the PAQ and the AIQ. Adaptive scaffolding (mean) was added to reduce the measurement

error if both questionnaires measured other aspects of adaptive scaffolding or if one of the questionnaires was made less adequately, honestly, or seriously.

For hypothesis three, three linear regression analyses were executed to test if a linear relation existed between TJA based on student names and respectively adaptive scaffolding (PAQ), adaptive scaffolding (AIQ), and adaptive scaffolding (mean). For hypothesis four, three linear regression analyses were executed to test if a linear relation existed between TJA based on student names in combination with cause-and-effect relations and respectively adaptive scaffolding (PAQ), adaptive scaffolding (AIQ), and adaptive scaffolding (mean).

Results

Teachers' Judgment Accuracy

For the first research question, the population consisted of 78.57% of the whole population: 209 students (128 girls, 81 boys); condition one consisted of 70 students, condition two consisted of 72 students, and condition three consisted of 67 students.

To assess teachers' judgment accuracy, I calculated gamma correlations between teachers' judgments of student performances and students' performances on tests for the three conditions. The mean judgment accuracies were positive for the three conditions (see fig. 5. Figure six shows that teachers' judgment accuracy is the highest for condition two and the lowest for condition three (as expected).

By exploring the data, it was found that the assumption of normality is violated, because the Shapiro-Wilk statistic is significant for each condition ($\alpha = .05$), $p < 001$. Because a one-way between groups ANOVA is quite robust with respect to violations of the assumption of normality, a one-way between groups ANOVA is used to explore if there are

differences between TJA based on the three judgment conditions. The assumption of homogeneity of variance has not been violated, because Levene's statistic is not significant at $\alpha = .05$, F (2, 206) = .207, p = .387.

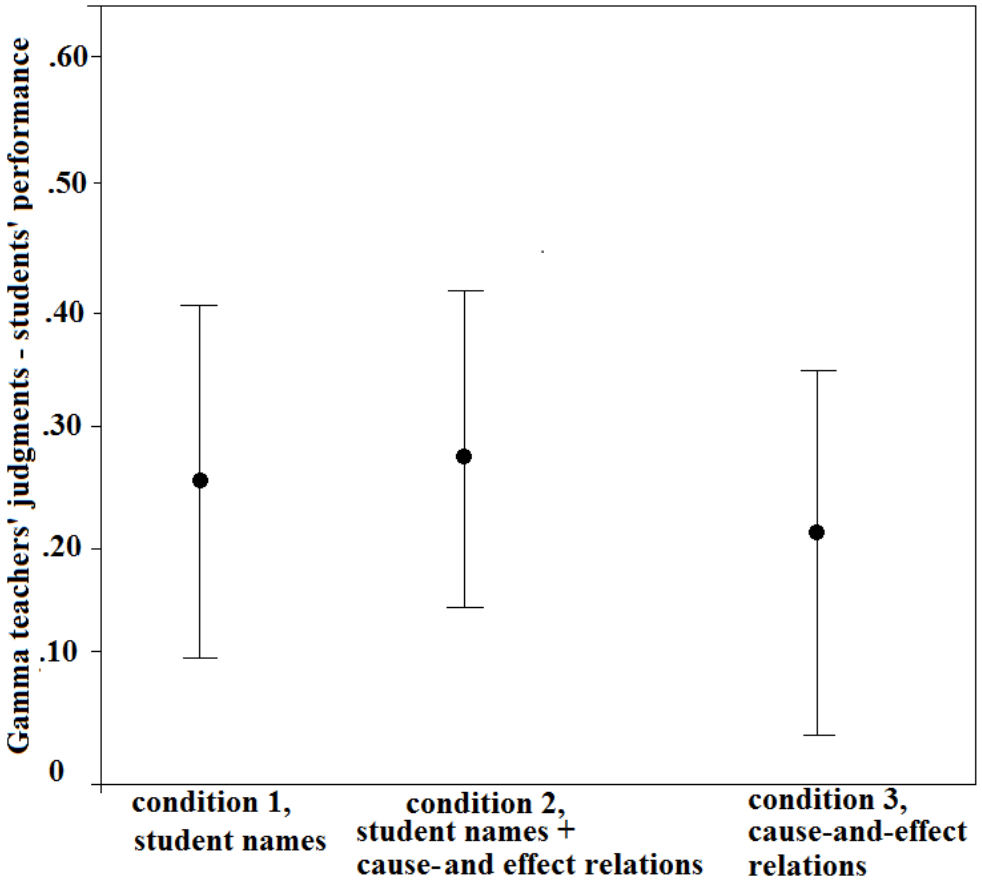

Figure 5. Teachers' judgment accuracy for each condition. 95% confidence interval is indicated by the error bars.

The ANOVA indicated that TJA did not significantly differ between the three judgment conditions, F (2, 206) = .207, p = .813, η^2 = .002.

Thus, both hypothesis one, which stated that TJA based on student names would be higher than TJA based on cause-and-effect relations, and hypothesis two, which stated that TJA based on student names in combination with cause-and-effect relations would be the highest among all three judgment conditions, were not confirmed. For the three comparisons

the effect sizes were respectively d = 0.03, 0.03, 0.06, indicating very small effects, based on Cohen's considerations of effect sizes (Cohen, 1988).

Adaptive Scaffolding

For research question two the amount of students participating decreased (see table 1).

Table 1.

The amount of students who filled in the questionnaires regarding adaptive scaffolding and the missing students

TJA condition	Filled in questionnaire(s)	Students	Missing students (%)
One	PAQ	41	41.43
	AIQ	46	34.29
	PAQ + AIQ	40	42.86
Two	PAQ	41	43.06
	AIQ	46	36.11
	PAQ+AIQ	39	45.83

Note. TJA condition one = TJA based on student names; TJA condition two = TJA based on student names in combination with cause-and-effect relations; Filled in questionnaire(s) = the questionnaires students filled in; Students = the amount of students who filled in the questionnaire(s); Participants (%) = the percentage students from the condition one (N = 70) and two (N = 72) who did not fill in the questionnaire(s).

Six simple linear regression analyses were executed, to estimate the proportion of variance in adaptive scaffolding by distinctly TJA based on student names and TJA based on student names in combination with cause-and-effect relations in adaptive scaffolding (PAQ), adaptive scaffolding (AIQ), and adaptive scaffolding (mean).

Before executing the analyses, some assumptions were assessed. The first assumption was met: all variables were at continuous level. The second assumption regarding normality

($\alpha = .05$) could not be met for respectively TJA based on student names, TJA based on student names in combination with cause-and-effect relations, and adaptive intervention (AIQ), $p < .001$, $p < .001$, $p = .004$, and could be met for respectively adaptive scaffolding (PAQ) and adaptive scaffolding (mean), $p = .597$, $p = .119$. No outliers were found when scores with > 3.29 SD from the mean were indicated (assumption three). The fourth assumption of linearity was assessed by making scatterplots. The scatterplots indicated no clear linear relations between respectively TJA based on student names and TJA based on student names in combination with cause-and-effect relations on adaptive scaffolding (PAQ), adaptive scaffolding (AIQ), and adaptive scaffolding (mean). Because of the violated assumptions, the intention was to execute non-linear regression analyses. Curve estimation explored which curve caused a model with a good fit, so the parameters could have been used for a non-linear regression. Unfortunately, for quadratic, cubic, compound, growth, exponential, and logistic, the models were non-significant and had a poor fit between respectively TJA based on student names and TJA based on student names in combination with cause-and-effect relations, and adaptive scaffolding (PAQ), adaptive scaffolding (AIQ), and adaptive scaffolding (mean) ($\alpha = .05$). Because no specified function could have been made that describes the relationship between the dependent variable and independent variable, a non-linear regression analysis could not be executed. For this reason, six linear regression analyses were performed.

Performing single linear regression analyses revealed that distinctly TJA based on student names and TJA based on student names in combination with cause-and-effect relations did not explain significant proportion of the variability ($\alpha = .05$) in respectively adaptive scaffolding (PAQ), adaptive scaffolding (AIQ), and adaptive scaffolding (mean) (see table 2).

Table 2

Proportion variance explained in adaptive scaffolding by TJA in condition one and two

TJA condition	Dependent variable	Proportion explained variance (%)	R^2	Adjusted R^2	F	p
One	PAQ	0.4	.00	-.022	(1, 39) = .152	.689
	AIQ	5.7	.06	.036	(1, 44) = 2.67	.110
	Mean	2.6	.03	.001	(1, 38) = 1.03	.318
Two	PAQ	3.2	.03	.007	(1, 39) = 1.29	.263
	AIQ	5.9	.06	.037	(1, 44) = 2.75	.104
	Mean	8	.08	.055	(1, 37) = 3.22	.081

Note. TJA condition one = TJA based on student names; TJA condition two = TJA based on student names in combination with cause-and-effect relations; Dependent variable PAQ = adaptive scaffolding measured by the PAQ; Dependent variable AIQ = adaptive scaffolding measures by the AIQ; Dependent variable mean = adaptive scaffolding measured by the mean of the scores on the PAQ and AIQ.

Thus, hypothesis three and four, which stated that TJA based on student names and TJA based on student names in combination with cause-and-effect relations predicted adaptive scaffolding, were not confirmed for the three different measures of adaptive scaffolding.

Discussion

The current study's main goal was to examine if TJA could predict adaptive scaffolding. Before examining this, the study investigated if TJA differed between teachers' judgments of students' performances in reading comprehension based on student names, student names in combination with by students generated cause-and-effect relations, and cause-and-effect relations. The results indicated two findings.

The study first found that TJA did not differ between teachers' judgments of students' performances in reading comprehension based on student names, on student names in combination with by students generated cause-and-effect relations, and on cause-and-effect relations. TJA based on student names was slightly higher than TJA based on by students generated cause-and-effect relations, but not significantly higher. This result is not in line with the reasoning that was based on findings from Hoth et al. (2016) and Thiede et al. (2016). Note that the means for TJA based on student names, student names in combination with cause-and-effect relations, and cause-and-effect relations were not high, respectively *M* = .26, .28, and .21. Because Koriat (1997) suggests that the higher cue diagnosticity, the higher TJA, it is possible that when teachers base their judgments on student names, they use cues with the same low cue diagnosticity as when teachers base their judgments on cause-and-effect relations that causes low TJA for both conditions. Note that the majority of the student participants were girls. Girls are better at reading comprehension than boys (Leinhardt, Seewald, & Engel, 1979; Logan & Johnston, 2009). Hoge & Butcher (1984) suggest that students' gender does not influence teachers' judgments. It is possible that the teachers in the current sample did not use the cue regarding better reading comprehension in girls than in boys. It might be that if they had used this diagnostic cue, TJA based on student names would have been significantly higher than TJA based on cause-and-effect relations.

TJA based on student names in combination with cause-and-effect relations resulted in the highest TJA among the three conditions, but not significantly higher. I expected TJA based on student names in combination with cause-and-effect relations to be significantly higher than the other two conditions, because teachers could make use of diagnostic cues from both conditions, resulting in the highest cue diagnosticity. This finding is not in line with Koriat (1997), who suggests that the higher cue diagnosticity, the higher the accuracy. A

possible explanation is that when offering student names in combination with cause-and-effect relations, teachers base their judgments on their preferred perspective. Hoth et al. (2016) found that not every teacher who has the availability to use a student-perspective in combination with a content-related perspective really uses both perspectives. Teachers who have more pedagogical knowledge and less content knowledge, prefer the use of a student-related perspective. Contrarily, teachers who have more content knowledge and less pedagogical knowledge, prefer the use of a content-related perspective. Some teachers use both perspectives (Hoth et al., 2016). This might explain why TJA based on student names in combination with cause-and-effect relations was not significantly higher than the other judgment conditions.

The second finding is that TJA based on both conditions did not predict more variance in adaptive scaffolding (in PAQ, AIQ, and mean) than had been expected by chance. This is not in line with Thiede et al. (2015) and Helmke and Schrader (1987), who suggested that TJA is associated with teachers' effectivity to adapt their instruction to the students.

For both TJA based on student names and TJA based on student names in combination with cause-and-effect relations more variance could be explained in adaptive scaffolding (AIQ) than in adaptive scaffolding (PAQ). This resulted from the higher standard deviation from the mean in the scores on the AIQ (SD = 15.63) than in the scores on the PAQ (SD = 8.65).

For TJA based on student names in combination with cause-and-effect relations, the highest proportion variance explained in adaptive scaffolding was found in adaptive scaffolding (mean) (almost significant), which contains the relatively high proportions variance explained in adaptive scaffolding (PAQ) and adaptive scaffolding (AIQ). However,

for TJA based on student names, the proportion variance explained in adaptive scaffolding (mean) was the lowest. An explanation for this is that the means from the relatively very low proportion of variance explained in adaptive scaffolding (PAQ) and the relatively high proportion of variance explained in adaptive scaffolding (AIQ) caused a lower proportion of variance in adaptive scaffolding (mean) than in adaptive scaffolding (AIQ).

A possible explanation for the non-significant findings on hypotheses three and four are missing data in condition one and two (see table 1). 30 to 40 percent of the students did not complete the questionnaires and some of the participating students only partly completed the questionnaires. It could be that the students were tired of the whole procedure and did not fill in the questionnaires seriously and honestly. This could reduce the internal validity, by not actually measuring the intended variable—adaptive scaffolding. This could also explain why different mean scaled scores on the PAQ and the AIQ were found. Another reason for different scores on the PAQ and the AIQ is that they measure different aspects of adaptive scaffolding. A second possible explanation for the non-significant findings is the non-existence of a relation between TJA and adaptive scaffolding.

Strengths of the Study

The study has five strengths. The first two strengths rely on internal validity—specifically methodological validity—, the third and fourth strength rely on external validity, and the last strength rely on the reliability of the study. Firstly, the students were randomly chosen as 'focus-students'. This ensured that all students had an equal opportunity to end up in the sample. Secondly, internal validity was increased by triangulation; two questionnaires were used to measure adaptive scaffolding. Thirdly, the study has the strength that students were derived from different regions of the Netherlands, which enhanced generalizability of

the results for students in the Netherlands. Fourthly, the teachers taught different subjects, which enhanced generalizability of the results for teachers who teach different subjects. Finally, the inter-rater reliability was good (Kappa = .62) (Cohen, 1988). A good inter-rater reliability indicates that personal characteristics do not influences the use of the instrument.

Limitations of the Study

The study has four limitations. Firstly, it remains unknown if the results are representative for students from lower or higher HAVO or VWO grades, pre-vocational secondary education (VMBO; four-year duration), pupils from primary schools, or students from countries with other educational systems. The test has been carried out on the ninth grades of HAVO or VWO, because the materials were developed for these particular levels. Because of the possibility that the results are not generalizable, the external validity decreases. Secondly, the current study did no investigation on the cues used by teachers and their diagnosticity for students' performances. Thirdly, because of the high percentages of missing data, the sample size decreased and might have been a cause for non-significant results. Moreover, it is uncertain if the data from the questionnaires are valid, because students' concentration levels were low at the end of the experiment, so several students might not have filled in the questionnaires seriously and honestly. This could reduce internal validity, by not really measuring adaptive scaffolding. Finally, the current study did not explore the reason for the different mean scores on the PAQ and the AIQ. The limitations from the current study are used for recommendations for future research.

Importance of the Results and Practical Implications

Despite the limitations, the results of the study are of interest. The results indicated no significant effects, but the direction of the effects were as expected. As expected, TJA based

on student names in combination with cause-and-effect relations caused the highest TJA, followed by TJA based on student names and TJA based on cause-and-effect relations. Besides this, TJA based on student names in combination with cause-and-effect relations predicted more variance in adaptive scaffolding than TJA based on student names, but did not explain more variance in adaptive scaffolding than had been expected by chance. If the reason for no significant differences lies in a sample size that was not large enough to produce significant differences, but the direction of the effects remains the same, the implications for educational practice are as follows. Teachers must know that a student-related judgment perspective could result in a higher TJA than a content-related judgment perspective, and the use of both of perspectives could result in the highest TJA. Teachers with a content-related judgment perspective should develop pedagogical knowledge, and teachers with a student-related judgment perspective should develop content knowledge. By doing this, they could benefit from the use of both perspectives. Besides this, teachers must know that, when their judgments are based on student names in combination with cause-and-effect relations, the accuracy of these judgments might predict their adaptive scaffolding as perceived by students.

Suggestions for Future Research

Three recommendations for future research were made. Firstly, future research is needed to examine if the results of the current study could be generalized. Future research could adapt their method and instruments to other educational levels. Secondly, future research could examine which cues are available when teachers make judgments of students' performances based on cause-and-effect relations. By doing this, diagnosticity of these cues for students' performances could be examined. After this, research could examine which cues

teachers actually use when making these judgments. This could be executed by making thinking aloud protocols (TAPs), in which teachers tell aloud what they think when making judgments. Besides this, for judgments based on student names, it is unknown which particular cues teacher really use from the available cues to make judgments. By doing this, it could be examined if the cues teachers use correspond to the cues with the highest diagnosticity for predicting students' performances. Because of the low mean TJA in the current research, it is expected that teachers have used cues with low diagnosticity. If future research also reveals low mean TJA when TJA is based on the three conditions from the current research, it is relevant to examine if they really used cues with low diagnosticity. It is also relevant to explore if teachers' use of a student-related or a content-related judgment perspective cause the use of cues with low diagnosticity. If this is the case, teachers could learn which cues to use with higher cue diagnosticity to enhance TJA. Finally, future research could enhance students' concentration for completing questionnaires, for example by spreading the data collection over more than one day. By this, internal validity is enhanced by actually measuring adaptive scaffolding. If students have the same scaled mean scores on the PAQ and the AIQ, there is more evidence that students in the current study did not fill in the questions seriously or honestly. If students scaled mean scores on the PAQ and the AIQ are significantly different from each other, it is possible that both questionnaires measure different aspects of adaptive scaffolding.

Conclusion

This study did not find any significant differences in the accuracy of teachers' judgments of students' performances in reading comprehension between judgments based on student names, student names in combination with by students generated cause-and-effect

relations, and cause-and-effect relations only. The results indicate that teachers' judgment accuracy did not predict adaptive scaffolding. Future research is necessary to address the limitations of the current research. It is important to gain more insight in methods to enhance TJA, because educational practices are adjusted on teachers' judgments. Besides this, it is important to actually measure the construct adaptive scaffolding within a larger sample to assess if TJA actually does not predict adaptive scaffolding.

References

Alvidrez, J., & Weinstein, R. S. (1999). Early teacher perceptions and later student academic achievement. *Journal of Educational Psychology, 91*(4), 731-746. doi:10.1037/0022-0663.91.4.731

Anderson, M. C. M., Thiede, K. W. (2007). Why do delayed summaries improve metacomprehension accuracy? *Acta Psychologica, 128*, 110-118. doi:10.1016/j.actps-y.2007.1 0.006

Borko, H., & Niles, J. (1982). Factors contributing to teachers judgments about students and decisions about grouping students for reading-instruction. *Journal of Reading Behavior, 14*, 127-140. doi:10.1080/10862968209547441

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. University of California Press.

Bürgermeister, A., Kampa, M., Rakoczy, K., Harks, B., Besser, M., Klieme, E., … & Leiβ, D. (2011). Dokumentation der befragungsinstrumente der interventionsstudie im projekt 'Conditions and consequences of classroom assessment' (Co^2CA). Retrieved from: www.pedocs.de

Butcher, K. R. (2006). Learning from text with diagrams: Promoting mental model development and inference generation. *Journal of Educational Psychology, 98*, 182-197. doi:10.1037/0022-0663.98.1.182

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2^{nd} ed.). Hillsdale, NJ: Erlbaum

Demaray, M. K., & Elliott, S. N. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performances. *School Psychology Quarterly, 13,* 8-24. doi:10.1037/h0088969

Diakidoy, I., Mouskounti, T., & Ioannides, C. (2011). Comprehension and learning from refutation and expository texts. *Reading Research Quarterly, 46,* 22-38. doi:10.1598/-RRQ.46.1.2

Dunlosky, J., & Thiede, K. W. (2013). Metamemory. In D. Reisberg (Ed.), Oxford handbook of cognitive psychology (pp. 283–298). New York: Oxford University Press. doi:10.-1093/oxfordhb/9780195376746.013.0019

Feinberg, A. B., & Shapiro, E. S. (2003). Accuracy of teacher judgments in predicting oral reading fluency. *School Psychology Quarterly, 18,* 52-65. doi:10.1521/scpq.18.1.52.2-0876

Feinberg, A. B., & Shapiro E. S. (2009). Teacher accuracy: An examination of teacher-based judgements of students reading with differing achievement levels. *Journal of Educational Research, 102,* 453-462. doi:10.3200/JOER.102.6.453-462

Fernández, M., Wegerif, R., Mercer, N., & Rojas-Drummond, S. (2001). Re-conceptualizing 'scaffolding' and the zone of proximal development in the context of symmetrical collaborative learning. *The Journal of Classroom Interaction, 36,* 40-54. Retrieved from: https://jstor.org/stable/23869224

Gaines, M. L., & Davis, M. (1990). *Accuracy of Teacher Prediction of Elementary Student Achievement.* Received from: ERIC database. (ED320942)

George, D., & Mallery, P. (2003). SPSS for Windows step by step: A simple guide and reference (4th ed.) Boston: Allyn & Bacon

Gobert, J. D., & Clement, J. J. (1999). Effects of student-generated diagrams versus student-generated summaries on conceptual understanding of causal and dynamic knowledge in plate tectonics. *Journal of Research in Science Teaching, 36,* 39-53. doi:10.1002//-(sici)1098-2736(199901)36:1<39::AID-TEA4?3.0.CO;2-1

Guay, F., Marsh, H. W., & Boivin, M. (2003). Academic self-concept and academic achievement: Developmental perspectives on their causal ordering. *Journal of educational psychology, 95*(1), 124-136. doi:10.1037/0022-0663.95.1.124

Helmke, A., & Schrader, F. W. (1987). Interactional effects of instructional quality and teacher judgment accuracy on achievement. *Teaching & Teacher Education, 3,* 91-98. doi:10.1016/0 742-051X(87)90010-2

Hoge, R. D., & Butcher, R. (1984). Analysis of teacher judgments of pupil achievement levels. *Journal of Educational Psychology, 78,* 777-781. doi:10.1037/0022-0663.76.5.777

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of literature. *Review of Educational Research, 59,* 297-313. doi:10.3102/0034-65430590 03297

Hoth, J., Döhrmann, M., Kaiser, G., Busse, A, König, J., & Blömeke, S. (2016). Diagnostic competence of primary school mathematics teachers during classroom situations. *ZDM Mathematics Education, 48,* 41-53. doi:10.1007/s11858-016-0759-y

Kaiser, J., Möller, J., Helm, F., & Kunter, M. (2015). The student inventory: How student characteristics bias teacher judgments. *Zeitschrift fur Erziehungswissenschaft, 18*, 279-302. doi:10.1007/s11618-015-0619-5

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*, 349-370. doi:10.1 037/0096-3345.126.4.349

Leinhardt, G., Seewald, A. M., Engel, M. (1979). Learning what's taught: Sex differences in instruction. *Journal of Educational psychology, 71*(4), 432-439. doi:10.1037/0022-0663.71.4.432

Leopold, C., & Leutner, D. (2012). Science text comprehension: Drawing, mean idea selection, and summarizing as learning strategies. *Learning and Instruction, 22*, 16-26. doi:10.1016/j.learninstruc.2011.05.005

Logan, S. & Johnston, R. (2009). Gender differences in reading ability and attitudes: examining where these differences lie. *Journal of Research in Reading, 32*, 199-214. doi:10.1111/j.1467-9817.2008.01389.x

Maki, R. H., & Serra, M. (1992). Role of practice tests in the accuracy of test predictions on text material. *Journal of Educational Psychology, 84*(2), 200-210. doi:10.1037/0022-0663.84.2.200

McCrudden, M. T., Schraw, G., Lehman, S., & Poliquin, A. (2007). The effect of causal diagrams on text learning. *Contemporary Educational Psychology, 32*, 367-388. doi:10.1016/j.cedpsych.2 005.11.002

Mega, C., Ronconi, L., & De Beni, R. (2014). What makes a good student? How emotions, self-regulated learning, and motivation contribute to academic achievement. *Journal of Educational Psychology, 106,* 121-131. doi:10.1037/a0033546

Molenaar, I., & Roda, C. (2008). Attention management for dynamic and adaptive scaffolding. *Pragmatics & Cognition, 16,* 224-271. doi:10.1075/pc.16.2.04mol

Pyc, M. A., Rawson, K. A., & Aschenbrenner, A. J. (2014). Metacognitive monitoring during criterion learning: When and why are judgments accurate? *Memory & Cognition, 42*, 886-897. doi:10.3758/s13421-014-0403-4

Schiefele, U. (1992) Topic interest and levels of text comprehension. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Seaton, M., Parker, P., Marsh, H. W., Craven, R. G., & Yeung, A. S. (2013). The reciprocal relations between self-concept, motivation, and achievement: juxtaposing academic self-concept and achievement goal orientations for mathematics success. *Educational Psychology, 34*(1), 49-72. doi:10.1080/01442310.2013.825232

Shavelson, R. J., & Borko, H. (1979). Research on teachers' decisions in planning instruction. *Educational Horizons, 57,* 183-189. Retrieved from: http://jstor.org/stable/42924342

Shavelson, R. J., & Stern, P. (1981). Research on teachers' pedagogical thoughts, judgments, decisions, and behavior. *Review of Educational Research, 51,* 455-498. doi:10.2307/-1170362

Thiede, K. W., Anderson, M. C., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66-73. doi:10.10-37/0022-0663.95.1.66

Thiede, K. W., Brendefur, J. L., Osguthorpe, R. D., Carney, M. B., Bremner, A., Strother, S., ... & Jesse, D. (2015). Can teachers accurately predict student performance? *Teaching and Teacher Education, 49*, 36-44. doi:10.1016/j.tate.2015.01.012

Van de Pol, J., & Elbers, E. (2013). Scaffolding student learning: A micro-analysis of teacher-student interaction. *Learning, Culture and Social Interaction, 2*(1), 32-41. doi:10.1016/j.lcsi.2012.12.001

Van de Pol, J., Kroonenberg, M., Poorthuis, A., & Mainhard, T. (In progress). Construction and validation of the perceived adaptivity questionnaire.

Van der Stuyf, R. R. (2002). Scaffolding as a teaching strategy. *Adolescent Learning and Development, 52*, 5-18. Received from: https://workplacesafety.pbworks.com

Van Loon, M. H., De Bruin, A. B., Van Gog, T., Van Merriënboer, J. J., & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy. *Acta Psychologica, 151*, 143-154. doi:10.1016/j.actpsy.2 014.06.007E

Vygotsky, L.S. (1978). *Mind in Society. The Development of Higher Psychological Processes.* London: Harvard University Press, Cambridge, Mass.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91*, 461-481. doi:10.1037/0033-2909.91.3.461

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem-solving. *Journal of Child Psychology and Psychiatry, and Allied Disciplines, 17*, 89-100. doi:10.1111-/j.1469-7610.1976.tb00381.x.

Wood, D., Wood, H., & Middleton, D. (1978). An experimental evaluation of four face-to-face teaching strategies. *International Journal of Behavioral Development, 1*, 131-147. doi:10.1177/016502547800100203

Appendix A

Text 'Using Botox'

Botox is the abbreviation of botulinum toxin; it is a poison produced by the bacterium Clostridium botulinum. This drug blocks the signal between the nerves in the skin muscles. The use of the drug is allowed since 1989. There is strict control in the Netherlands. In America, there has been an accident in 2004 with the dosage of Botox, in which 28 people died. Due to blocking of the signal between the nerves and the skin, Botox was originally used against muscle contractions, for example with patients who continuously blinked their eyes. Because Botox blocks the signal between the nerves and the skin muscles, this drug is also used in plastic surgery to smoothen the wrinkles next to the eyes and on the forehead. The effect of a treatment lasts between one and six months. The treatment makes people look younger. However, because the wrinkles between the eyes and on the forehead disappear, someone's face expression could undesirably change (Van Loon et al., 2014).

Text 'Music makes smarter'

Learning to play an instrument could have many advantages. This is because for reading sheet music, one must make use of different regions of the brains. The brains connect music with memory in various brain regions such as the amygdala and the hippocampus. Canadian research showed that as a result of the use of these different brain areas, 12-year-old children who learned to play music scored higher on an IQ test. Also, the use of brains during learning music could enhance spatial skills that are useful in solving mathematical tasks such as fractions. Besides that, music can help with the retrieval of memories. For example, an old song reminds you of something that happened long ago. Also, for example patients with Alzheimer's disease are helped by music to retrieve memories. This was demonstrated in

research of scientist Polk. He tested an old woman with Alzheimer's, who seemed to recognize all the music she learned in the past (Van Loon et al., 2014).

Appendix B

Distribution of tasks

Theoretisch kader, probleemstelling, onderzoeksvragen

Deadline: vrijdag 11 maart 2016

Anne-leen Boeve:

Opzoeken van literatuur voor de theoretische inleiding over:

- Accuratesse van beoordelingen van leerkrachten op het gebied van lezen
- Het genereren van oorzaak-en-gevolg relaties
- Contingentie van instructie, een voorwaarde voor scaffolding

Gevonden literatuur beknopt in schema's beschrijven van introductie en resultaten.

Schrijven van theoretische inleiding over:

- Accuratesse van beoordelingen van leerkrachten

Dagmar Bos:

Opzoeken van literatuur voor theoretische inleiding over:

- Accuratesse van beoordelingen van leerkrachten op het gebied van lezen
- Het genereren van oorzaak-en-gevolg relaties

Sylvia Kasperink:

Opzoeken van literatuur en schrijven theoretische inleiding over:

- Introductie (wetenschappelijke relevantie)
- Accuratesse van beoordelingen van leerkrachten op het gebied van lezen
- Het genereren van oorzaak-en-gevolg relaties
- Zone van Naaste Ontwikkeling en adaptieve scaffolding
- Contingentie van instructie, een voorwaarde voor scaffolding

- De huidige studie

Deadline: dinsdag 15 maart 2016

Samen:

Controleren theoretische inleiding: samenhang, spelling, overgangen etc.

Methode:

Deadline: vrijdag 18 maart 2016

Anne-leen Boeve:

Schrijven van:

- 6.2.2. Instrumenten
- 6.2.3. Design en procedure

Dagmar Bos:

Schrijven van:

- 6.2.1. Deelnemers
- 6.2.2. Instrumenten
- 6.2.3. Design en procedure

Het werven van vijf klassen op één school voor het onderzoek

Sylvia Kasperink:

Schrijven van:

- 6.2.2. Instrumenten
- 6.2.3. Design en procedure

Deadline: zondag 21 maart 2016

Samen:

Controleren

Deadline: zaterdag 19 maart 2016

Samen:

Nadenken over analysetechnieken

Schrijven van:

- 6.2.4. Analyse

Toegevoegde waarde:

Deadline: donderdag 17 maart 2016

Anne-leen Boeve en Dagmar Bos:

Schrijven van:

- 6.5. Toegevoegde waarde praktisch belang

Sylvia Kasperink:

Schrijven van:

- 6.4. Wetenschappelijk belang

Deadline: zondag 20 maart

Samen:

Controleren samenhang, spelling, overgangen etc.

Planning en risico's:

Deadline: zondag 20 maart 2016

Sylvia Kasperink:

Schrijven van:

- 7.1. Gedetailleerde planning in weken

Samen:

Nadenken over risico's

Schrijven van: 7.2. Risico's

Samenvatting:

Deadline: zondag 20 maart 2016

Sylvia Kasperink:

Schrijven van:

- 5. Titel en samenvatting

Deadline: maandag 21 maart 2016

Samen:

Controleren van de samenvatting

Eindcontrole

Deadline: maandag 21 maart 2016

Samen

Het onderzoeksplan controleren

Inleveren concept onderzoeksplan

Deadline: donderdag 24 maart 2016

Samen

Het onderzoeksplan inleveren

Verbeteren concept onderzoeksplan

Deadline: woensdag 5 april 2016

Samen

Het onderzoeksplan verbeteren

Inleveren definitief onderzoeksplan

Deadline: donderdag 7 april 2016

Samen

Het onderzoeksplan verbeteren

Sorteren materialen dataverzameling

Datum: 14 april 2016

Samen

Materialen voor dataverzameling sorteren voor ieder groepje (op school)

Deadline: 18 april

Dagmar Bos

Materialen voor dataverzameling sorteren op klas en leerkracht (thuis)

Werven docent en klas andere school (Zwolle)

Deadline: 16 april 2016

Sylvia

Andere docent en klas op een andere school dan in Houten werven, vanwege afzegging één docent met bijbehorende klas in Houten

Data verzamelen Zwolle

Datum: 19 april 2016

Samen

Data verzamelen

Data verzamelen Houten

Datum: 20 april 2016

Anne-leen Boeve en Dagmar Bos

Data verzamelen

Scoren diagrammen en scores invoeren

Deadline: 16 mei 2016

Dagmar Bos

Scoren van diagrammen

Invoeren scores

Invoeren algemene informatie leerkracht en leerlingen

Invoeren responsen van vragenlijsten

Deadline: 24 mei 2016

Anne-leen Boeve en Sylvia Kasperink

Invoeren responsen van vragenlijsten in SPSS

Samenvoegen van het databestand met algemene informatie leerkrachten en leerlingen, scores op diagrammen en het databestand met responsen op vragenlijsten in SPSS

Inleveren concept bachelor thesis

Deadline: 30 mei 2016

Individueel

Individueel inleveren concepten bachelor thesis

Verbeteren concept bachelor thesis

Deadline: 7 juni 2016

Individueel

Individueel verbeteren concepten bachelor thesis

Inleveren definitieve bachelor thesis

Deadline: 8 juni 2016

Individueel

Individueel inleveren definitieve bachelor thesis

Voorbereiden presentatie congresdag

Deadline: 8 juni 2016

Samen

Vergelijken thesis en resultaten

Verklaringen resultaten vergelijken

Voorbereiden presentatie thesis

Geven presentatie congresdag

Datum: 9 juni 2016

Samen

Geven van presentatie op de congresdag